Journal of Informatics and Web Engineering

Vol 1 No 1 (2022) eISSN: 2821-370X

Performance of Sentiment Classifiers on Tweets of Different Clothing Brands

Muhammad Shafiq[1], Hu Ng[2*], Timothy Tzen Vun Yap[3], Vik Tor Goh[4]

[1,2,3] Faculty of Computing and Informatics, Multimedia University Cyberjaya, Malaysia.
[4] Faculty of Engineering, Multimedia University, Cyberjaya, Malaysia.
**Corresponding author: (nghu@mmu.edu.my)*

Abstract - Social media such as Facebook, Instagram, LinkedIn, and Twitter ease the sharing of ideas, thoughts, videos, and photos and information through the building of virtual networks and communities. This has allowed companies and products to reach a wider audience in terms of marketing and advertising, and to gauge feedback from the public. This research investigates clothing brand mentions on Twitter to perform sentiment analysis on users' thoughts on three clothing brands, namely Asos, Uniqlo and Topshop. The data is collected by applying python libraries, Tweepy to access data from the Twitter streaming API. Following that, data pre-processing such as tokenization, filtering, stemming, and case normalization are performed to remove outliers. Then, the TextBlob algorithm is applied to label the tweet data into three classes; Positive, Negative and Neutral based on the polarity of the tweets. Word embeddings are also created using Word2Vec with TF-IDF. The word embeddings are fed into classification models namely Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF), Logistic Regression (LR) and Multilayer Perceptron (MLP) by comparing their accuracy performances. The models went through training and testing process on a curated tweet dataset comprising 24000 records with three clothing brands (Asos, Uniqlo, Topshop). The classification process was carried out by SVM, NB, RF, LR and MLP with a ratio of 50-50 and 70-30 train-test splits. Hyperparameter tuning was implemented by GridSearchCV to find the best parameters of classification models in order to optimize the best results. The evaluation of performance was measured with accuracy, precision, recall and F1-Score. In the 50-50 train-test splits, LR achieved the highest accuracy by scoring 82%, 87% and 87% on Asos, Uniqlo and Topshop respectively. In the 70-30 train-test splits, LR also achieved highest accuracy by scoring 85%, 90% and 90% for the three clothing brands respectively.

Keywords— Sentiment analysis, classification, machine learning, clothing brand

Received: 17 January 2022; Accepted: 6 March 2022; Published: 16 March 2022

I. INTRODUCTION

As the number of reviews of product rapidly grows through social media, analyzing it through natural language processing can be very beneficial and useful tool instead of looking through each review manually. Sentiment analysis is intended to define the sentiment and polarity of a portion of text [1]. Generally, language phrases are under two forms of statement, such as fact statement and a non-fact statement, Those statements are identified as objective and subjective in categorical phrases. Facts or objective phrases are related to events entities and their properties. In contrast, non-fact statement is typically related to a person's emotions, feelings, or thoughts.

Journal of Informatics and Web Engineering
https://doi.org/10.33093/jiwe.2022.1.1.2

Published by MMU Press. URL: https://journals.mmupress.com/jiwe

This project aims to investigates clothing brand mentions on Twitter to perform sentiment analysis on users' thoughts on three clothing brands, namely Asos, Uniqlo and Topshop. The proposed system can monitor the consumer reactions and feedback on new products that are freshly launched in the market. It is able to help companies to react on the feedbacks in a fast manner and take them as the reference for their future products. It can also help the consumers to compare various brands on similar products and determine the best choice of their preferences.

II. LITERATURE REVIEW

A. Related work

Azzouza et al [2] proposed a real-time application approach that is able to explore and track thoughts on Twitter by utilizing Apache Storm. The application offers various thoughts' representations by dynamic graphic visualizations and able to suggest appropriate keywords on the topic of same interest.

Paltoglou et al. [3] developed an instinctive, non-domain, unsupervised lexicon-based method, which is able to estimate the degree of emotional strength from the texts. The system can make prediction on keywords by implementing on two distinct but paired contexts: subjectivity discovery and polarity categorization.

Abdullah et al. [4] performed sentiment analysis from Netno-graphy data that were gathered from various social media platforms. They applied AYLIEN, Monkey-learn and Text Analysis API in order to obtain five classes of sentiment polarities, namely Positive, Negative, Sarcastic, Ideology and Neutral sentiments.

Fronzetti Colladon et al. [5] applied Semantic Brand Score to visualize 206,000 tweets for the mentions of the fashion brands Fendi, Gucci and Prada. They determined Gucci dominated the discourse, with highest values of prevalence, variety, and connectivity.

Yuan et al. [6] proposed a framework that combines three elements from images, posted texts and fashion attributes as modality to be encoded respectively, and then merged as a multimodal composer. They collected over 12k fashion related data from Instagram using a set of pre-defined hashtags to perform the sentiment analysis.

Liu et al. [7] visualized the consumers' clothing consumption progression during the COVID-19 global pandemic in 2020 from 68,511 relevant tweets. They incorporated the perspectives of lifestyles and tension handling to analyse consumers' reactions to clothing consumption.

Choi et al [8] performed semantic network analysis on tweets to find out the impact of social media, influencers, fashion brands, designers that were mentioned in all four major cities (London, Paris, Milan, and New York) during the 2019 Fall/Winter Fashion Week. Their work provided valuable data for allowing fashion retailers to improve their marketing tactics.

B. Sentiment classification and feature selection

Logistic Regression (LR) is a binary linear regression model that is mainly driven by the posterior probability of each class [9]. LR is a great starter algorithm for text related classification and is able to perform well even in an imbalanced dataset [10].

SVM is capable of developing the best possible border, a line known as a hyperplane to separate dimensional spaces into difference clusters or classes [11]. SVM will locate the hyperplane that will correctly differentiate those classes. The best hyperplane obtained will be the one that maximizes the margins from different classes.

Naïve Bayes (NB) is formed with the Bayes' rules of conditional probability and holds superior capabilities in performing well in a large dataset [12]. Based on the Bayes' rules, NB estimates the probability of a property by giving a set of records as verification. The posterior is calculated from the prior or likelihood of property and separable by its evidence.

Multilayer Perceptron (MLP), a form of Artificial Neutral Network (ANN) can provide better accuracy rate in a shorter time frame [13]. It applies a feedforward architecture by connecting the perceptron from the input to the output in one direction with several routes. Basically, it has 3 layers including input layer, output layer and one hidden layer. If it has more than 1 hidden layer, it is called a deep ANN.

Random Forest (RF) runs by forming a variety of decision trees during the training with certain parameters setting. It has been widely used in machine learning due to its capability to handle multiclass classification, resistance to outliers, smaller number of parameters to tune and accept embedded feature selection [14].

Feature selection is a tool for dimensionality reduction where it decreases the number of extracted features of a dataset in order to increase the classification performance. Redundant features in a dataset can defeat a machine learning classifier and might diminish its efficiency [15]. In the research work by Tang et al. [15] on a medical dataset, feature selection can successfully distinguish and enhance overall evaluation metrics.

III. RESEARCH METHODOLOGY

In this paper, the research work consists of five phases, namely Tweet data extraction, text labelling, data pre-processing, sentiment classification and confusion matrix. The flow of the processes is shown in Figure 1.

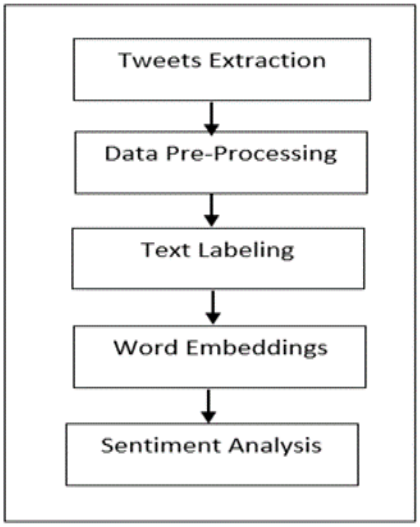

Figure 1. Phases of the study.

A. Tweet extraction

24,000 (8000 x three clothing brands) tweets were extracted by applying python libraries, Tweepy to access data from the Twitter streaming API. The three clothing brands are Asos, Uniqlo and Topshop.

B. Data pre-processing

Data pre-processing such as tokenization, filtering, stemming, and case normalization were performed to remove outliers and split the paragraph into a one-word window.

Tokenization is a process where text is split into several words. The definition of tokenization is that small tokens are created from large phrases or text to be further analyzed. For example, given a sentence “the man is wearing Topman shorts”. The expected output after tokenization will be ‘the’, ‘man’, ‘is’, ‘wearing’, ‘Topman’ and ‘shorts’.

In the case of normalization, the input were converted to lower case, so that misinterpretation could be minised during the classification process. In stemming, the first few letters or last few letters were sliced to reach the source term of the words. For instance, a word such as wearing is stemmed as wear only as the suffix and prefix are removed.

During the filtering process, stop words were removed. For instance, given a tweet of sentence ‘That is a Jordan 1’, ‘a’ and ‘is’ is a stop word. Upon going through the filtering process, stop word is removed and the sentence will later then be updated as ‘That Jordan 1’.

C. Text labelling

The TextBlob algorithm was applied to label the tweet data into three classes; Positive, Negative and Neutral based on the polarity of the tweets. Figure 2 shows the output from the labelling process.

	Tweets	Polarity	Subjectivity	Sentiment
0	sex scenes are bad john waynes politics the irishman is too long is marvel real cinema when jord...	-0.183333	0.455556	Negative
1	tarot of delphi 70€450€ 😐	0.000000	0.000000	Neutral
2	the only reason why I saved this is bc her husband is wearing a the devil wears prada shirt http...	0.000000	1.000000	Neutral
3	CHANYEOL in Prada was a delight\nEXO httpstcoyCuvXAhlpA	0.000000	0.000000	Neutral
4	PINK FLUFFLY PRADA BAG httpstcoUVGC5TVTII	-0.100000	0.300000	Negative
...	...	...	...	...
495	Wait Black Meryl as in Meryl Streep from The Devil Wears Prada 👀	-0.166667	0.433333	Negative
496	if exo does not have visuals how can exo members become ambassadors of Gucci Prada Dior Cartier ...	0.268750	0.443750	Positive
497	Just downloaded most of the albums by Erra WCAR and The Devil Wears Prada again	0.500000	0.500000	Positive
498	Great effort by American Magic to get back on the track but they were always going to be on t...	0.243229	0.406250	Positive
499	PINK FLUFFLY PRADA BAG httpstcoUVGC5TVTII	-0.100000	0.300000	Negative

Figure 2. Output from the labelling process.

D. Word embeddings

This research work utilized Word2Vec to perform word embedding. Word2Vec by Mikolov et al. [16] is a word embedding method that comprises of two structural models, namely Skip-gram and Continuous Bag-of-Words. As the Skip-gram structure has been shown better results in comparison to Continuous Bag-of-Words [16], this work utilized the Skip-Gram structure. Word embedding was performed in order to convert the tweets into floating-point numbers, which were stored in a high dimension array as a dictionary. Then, 300-dimension word vectors were generated from the high dimension array.

E. Design of Experiment

Five supervised machine learning models were implemented which were SVM, NB, RF, LR and MLP with 70-30 and 50-50 train-test splits. The experiments were evaluated with 10-fold cross validation. Experiments were carried out on all the three clothing brands (Asos, Uniqlo, Topshop), and each individual brand respectively. GridSearchCV was applied to perform hyperparameter tuning for results optimization. The performance metrics were accuracy, F1-Score, precision, and recall. Boruta feature selection algorithm was applied to select the features that positively contribute to the classification accuracy. The performance metrics were rate of accuracy, F1-Score, precision, and recall.

V. RESULTS AND DISCUSSIONS

The experiments results are shown in Tables 1 to 4.

Table 1. Performance evaluation on all three clothing brands with 70-30 and 50-50 train-test split.

	Train-test split ratio							
	70-30				50-50			
Classifier	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)
SVM	81	84	79	89	79	82	78	87
NB	69	73	71	76	66	71	69	73
RF	77	80	72	98	79	82	78	87
LR	79	82	77	88	84	87	81	94
MLP	76	79	76	82	74	76	74	78

Table 2. Performance evaluation on Topshop with 70-30 and 50-50 train-test split.

	Train-test split ratio							
	70-30				50-50			
Classifier	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)
SVM	82	85	81	89	80	83	79	89
NB	74	77	76	77	72	76	72	75
RF	83	85	81	88	81	84	79	87
LR	90	92	86	97	87	88	83	96
MLP	78	79	79	81	76	78	76	80

Table 3. Performance evaluation on Uniqlo with 70-30 and 50-50 train-test split.

	Train-test split ratio							
	70-30				50-50			
Classifier	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)
SVM	85	85	87	84	85	85	87	84
NB	76	76	86	71	73	73	83	68
RF	80	80	70	95	76	76	69	91
LR	90	90	91	91	87	87	86	90
MLP	84	84	89	81	81	81	85	77

Table 4. Performance evaluation on Asos with 70-30 and 50-50 train-test split.

	Train-test split ratio							
	70-30				50-50			
Classifier	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)
SVM	81	84	81	87	85	85	87	84
NB	69	74	72	76	68	73	70	75
RF	75	80	67	98	74	79	66	98
LR	85	88	83	94	82	85	81	91
MLP	74	77	80	74	72	73	69	78

In the experiments carried out on all the three clothing brands (Asos, Uniqlo, Topshop), LR scored the highest accuracy with a score of 84% on the total number of tweets in the 50-50 train-test split. Meanwhile, SVM with polynomial kernel scored the highest accuracy with a score of 81% on the total number of tweets data in the 70-30 train-test split.

In the experiments carried out on individual clothing brand, LR achieved the highest accuracy with scores of 82%, 87% and 87% on three individual clothing brands (Asos, Uniqlo and Topshop) respectively in 50-50 train-test split. LR also achieved highest accuracy with scores of 85%, 90% and 90% for the three clothing brands respectively in 70-30 train-test split.

It has been shown that the more data can be employed in the training, the higher the accuracy that can be achieved. The accuracy of the classifiers are shown to be higher in the classification for individual brand compared to having to classify the three clothing brands together.

VI. CONCLUSIONS

The paper compared the performance of sentiment classifiers on tweets of three clothing brands with five classifiers. As the study looks at only 20000 tweets of three clothing brands, more tweets will be included to determine the robustness of the classifiers in future.

ACKNOWLEDGEMENT

The authors received no funding from any party for the research and publication of this article.

REFERENCES

[1] C. Chauhan, and S. Sehgal, "Sentiment analysis on product reviews", IEEE International Conference on Computing, Communication and Automation (ICCCA), 2017, pp. 26-31.

[2] N. Azzouza, K. Akli-Astouati, A. Oussalah, and S.A. Bachir, "A real-time Twitter sentiment analysis using an unsupervised method", In Proceedings of the 7th International Conference on Web Intelligence, Mining and Semantics, 2017, pp. 1-10.

[3] G. Paltoglou, and M. Thelwall, "Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media", ACM Transactions on Intelligent Systems and Technology (TIST), vol. 3(4), pp. 1-19, 2012.

[4] N.S.D. Abdullah, and I.A. Zolkepli, "Sentiment analysis of online crowd input towards brand provocation in Facebook, Twitter, and Instagram", In Proceedings of the International Conference on Big Data and Internet of Thing, 2017, pp. 67-74.

[5] Fronzetti Colladon, A., Grippa, F., & Segneri, L. (2021, June). A new system for evaluating brand importance: A use case from the fashion industry. In *13th ACM Web Science Conference 2021* (pp. 132-136).

[6] Yuan, Y., & Lam, W. (2021). Sentiment Analysis of Fashion Related Posts in Social Media. *arXiv preprint arXiv:2111.07815*.

[7] Liu, C., Xia, S., & Lang, C. (2021). Clothing Consumption during the COVID-19 Pandemic: Evidence from mining tweets. Clothing and Textiles Research Journal, 39(4), 314-330.

[8] Choi, Y. H., Yoon, S., Xuan, B., Lee, S. Y. T., & Lee, K. H. (2021). Fashion informatics of the Big 4 Fashion Weeks using topic modeling and sentiment analysis. Fashion and Textiles, 8(1), 1-27.
[9] D.R. Cox, and E.J. Snell, Analysis of binary data, 2018.
[10] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey", *Information*, *10*(4), pp. 150, 2019.
[11] Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. Artificial Intelligence Review, 26(3), 159-190.
[12] A. Havan, and M. Harshil, "Student Performance Prediction using Machine Learning", International Journal of Engineering Research And V4(03), pp. 111–113, 2015. https://doi.org/10.17577/ijertv4is030127
[13] M.V. Amazona, & A.A Hernandez, "Modelling student performance using data mining techniques: Inputs for academic program development", ACM International Conference Proceeding Series, 2019, pp. 36–40. https://doi.org/10.1145/3330530.3330544
[14] R. Katuwal, P.N. Suganthan, and L. Zhang, "Heterogeneous oblique random forest", Pattern Recognition, 99, 107078, 2020.
[15] R. Tang, and X. Zhang, "CART Decision Tree Combined with Boruta Feature Selection for Medical Data Classification", 5th IEEE International Conference on Big Data Analytics, 2020, pp. 80–84. https://doi.org/10.1109/ICBDA49040.2020.9101199
[16] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality", In Advances in neural information processing systems, pp. 3111-3119, 2013.